SLC Walkthrough for Practice Set Week3

- Import required libraries.
- Load the data set.

Let's begin with some hands-on practice exercises:

1. Build a decision tree model using the gini criterion. And obtain the number of correctly predicted wheat seeds in the test set.

HINT:

Consider the train-test split in the ratio 70:30 with random state = 1.

- Split the dataset into Train and Test set with the above proportion level.
- Build Decision Tree Classifier with criterion = gini.
- Use confusion matrix to know the correctly predicted classes by the trained decision tree model.

2. Plot a decision tree for the model in the previous question and identify the seed type of the first observation in the test set.

HINT:

Import following libraries to plot the decision tree plot

from sklearn import tree import pydotplus

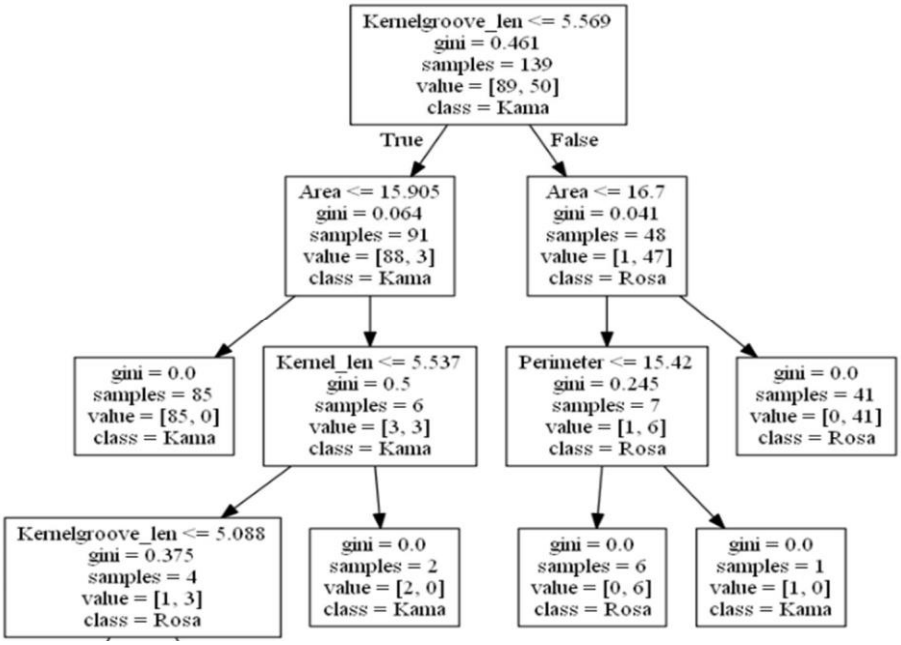
from IPython.display import Image

- Use Tree. export_graphviz and set the parameter appropriately.
  Ex: dot_data = tree.export_graphviz(decision_tree, feature_names = labels, class_names = ["Kama","Rosa"])

This file is meant for personal use by swethaviswanathan.r@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

- Plot the decision tree plot using DOT format in dot_data and store it in the variable lets say graph.
- Display the decision Tree using Image method.
  Ex: pass variable graph.create_png() into the Image method.

]:



- Interpret the results.

3. Select the optimal number for decision trees from the given list of values to build a random forest using entropy criterion.

Use the given list:

no_of_trees = [6, 8, 10, 12, 14, 16]

Consider the train and test set in Q1.

HINT:

- Use hyper parameter tuning for given params.
- Set criterion = entropy for the Random Forest model.
- Use Grid Search CV to tune the params.
- And get the best param from grid model.

This file is meant for personal use by swethaviswanathan.r@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

4. Identify the most important variable in the random forest build with the optimal number of trees obtained in Q3.

Consider the train and test set in Q1.

First build the random forest with the `entropy` criterion and optimal number of trees that you got in the above model.

HINT:

- Fit the Random Forest model with criterion = 'entropy' and n_estimators = value that you got in the Grid Search CV model.
- Use Random Forest feature importance method to identify the feature importance.

5. Which is the best criterion to build a decision tree for the given dataset?

HINT:

Use Grid Search CV to find the best criterion out of gini and entropy.

6. Build a random forest containing ten trees and compute the precision and sensitivity of the model from the confusion matrix.

HINT:

- Build Random Forest model with number of estimators is 10.
- Use confusion matrix to get the confusion matrix and from that calculate precision and recall.

7. Find the optimal depth of the decision tree from the given list of values.

Use the given list:

This file is meant for personal use by swethaviswanathan.r@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

depth_values = [3, 5, 7, 9]

HINT:

Use hyper parameter tuning using grid search cv to find the optimal number of maximum depth.

8. Build and plot a decision tree with maximum 5 terminal nodes (use the entropy criterion).

HINT:

Build decision tree model with criterion is entropy and maximum leaf node is 5.

Follow question 2 steps to plot the decision tree plot.

9. Build a random forest with entropy criterion such that each leaf node will contain at least three samples. Also calculate the f-1 score and accuracy of the model.

HINT:

- Build a model with following params, o Criterion is entropy.
    - o Minimum samples leaf is 3.
- Use classification report to get f1 score and accuracy.

10. Plot a decision tree with the optimal criterion such that it will contain no more than 4 terminal nodes and each terminal node will contain at least 5 observations.

HINT:

- Set the following params for decision tree.
    - o Criterion is entropy.

This file is meant for personal use by swethaviswanathan.r@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

- Maximum leaf nodes are 4.
- Minimum samples leaf is 5.

- Follow the question number 2 steps to plot the decision tree plot.

This file is meant for personal use by swethaviswanathan.r@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.